Stratified Nested Case-Control Sampling in the Cox Regression Model

Bryan Langholz
Department of Preventive Medicine,
University of Southern California, School of Medicine,
2025 Zonal Ave, Los Angeles, California 90033-9987, U.S.A.

Ørnulf Borgan
Institute of Mathematics, P.O. Box 1053 Blindern,
University of Oslo, N-0316 Oslo 3, Norway

20 August 1992

Abstract

A new type of nested case-control sampling is presented in which the sampled risk sets include the failure and random samples from strata defined by covariate information on all cohort subjects. This sampling may be non-representative in that the proportion sampled from each stratum need not be representative of the entire risk set. Asymptotic relative efficiency comparisons indicate that this type of sampling has superior efficiency to simple nested case-control sampling in situations of practical interest. A simple extension of the method is given which allows for non-representative sampling of failures. Analysis of stratified sampled data may be performed using standard conditional logistic likelihood software which allows for an "offset" in the model.

1 Introduction

Epidemiologic cohort studies of a rare disease require many subjects and/or long follow up periods in order to accumulate enough diseased subjects to have sufficient power to explore the variation in rates across various factors of interest. But, because they are large studies, collecting high quality covariate information on each subject is an expensive task. Thus, often only enough information is collected on the full cohort to allow one to draw a simple nested case-control sample (Thomas, 1977; Oakes, 1981) in which each diseased "case" is matched to a random sample of "controls" from those at risk at the case failure time. Additional detailed covariate information is then gathered for subjects in this sample to perform a proper analysis of the data. But, suppose there is additional information known for a significant portion of the cohort. For example, suppose exposure information has been gathered for all members of the cohort and it is desired to collect additional information on a sample of the cohort in order to assess the role of potential confounders or to study interactions of the exposure with other risk factors. Another possibility is that a crude

[0] *Key words and phrases.* Asymptotic efficiency; Cohort study; Case-control study; Design of medical study; Epidemiology; Martingale; Survival analysis

measure of exposure has been gathered for most of the cohort members and researchers wish to collect more precise exposure data on a much smaller sample of the cohort.

In drawing a nested case-control sample only the at-risk status of the cohort members is used. It is easy to imagine that incorporating other covariate information into the sampling process might lead to a more informative sample. This concept has been discussed in the context of grouped data situations and logistic regression in White (1982), Breslow and Cain (1988), Cain and Breslow (1988), Weinberg and Wacholder (1990) and Weinberg and Sandler (1991). Here, we develop methods for "stratified" sampling of controls in a modification of simple nested case-control sampling from a cohort. As it is often the situation that there are a relatively small number of cases, we will assume for the time being, that all cases are to be used and that sampling is of the risk sets at the failure times in the cohort. We discuss non-representive sampling of cases in Section 5.

We assume the Cox proportional hazards model (Cox, 1972) where the intensity for a subject with vector of covariates $Z(t)$ and censoring indicator $Y(t)$ at time t may be given as

$$\lambda(t) = Y(t)\lambda_0(t)\exp\{\beta_0' Z(t)\}. \tag{1}$$

Let t_j be the jth ordered failure time and i_j be the index of the failure at time t_j. At t_j a sampled risk set $\tilde{\mathcal{R}}(t_j)$ of size m, with $m-1$ controls and 1 case, is drawn as follows: Each person in the risk set, including the case, is classified into one of, say, L sampling strata. This classification cannot be based on case-control status. More precisely, if someone other than the actual case had been the case, it would not have resulted in a change in the way classification is done. Then $\tilde{\mathcal{R}}(t_j)$ is to consist of $m_l > 0$ subjects from the $n_l(t_j)$ at risk individuals in stratum l where the m_l are chosen in advance and do not need to reflect the representation of stratum l in the full risk set. When actually performing the sampling, one randomly samples, without replacement, m_l controls from stratum l except for the case's stratum from which one samples only $m_l - 1$ controls. The case is always included in the sample so that there are a total of m_l from stratum l. Let $A_k(t_j)$ be the sampling stratum for subject k at time t_j and $w_k(t_j) = n_{A_k(t_j)}(t_j)/m_{A_k(t_j)}$. Then, as will be shown in Section 3, the partial likelihood for the sampled data set is given by

$$\mathcal{L}(\beta) = \prod_{t_j}\left[\frac{\exp\{\beta' Z_{i_j}(t_j)\}w_{i_j}(t_j)}{\sum_{k\in\tilde{\mathcal{R}}(t_j)}\exp\{\beta' Z_k(t_j)\}w_k(t_j)}\right]. \tag{2}$$

This has the form of the usual "conditional logistic" likelihood where the contribution of a subject from stratum l is weighted by the inverse of the proportion in the sampled risk set from stratum l. The partial likelihood (2) has "basic likelihood properties" by which we mean that expectation of the score evaluated at β_0 equals zero, and the expected information matrix at β_0 equals the covariance matrix of the the score. Standard conditional logistic regression fitting algorithms may be used simply by including a subject's log weight, $\log w_i(t_j)$, as an "offset" in the model. This feature is currently available in many software packages.

As an example, consider a cohort in which a dichotomous exposure is known for all at risk subjects at each failure time. Additional information is to be collected on a sample and might include precise exposure measurements, confounder data, and other known or potential exposure information. In 1:1 matched stratified sampling, each exposed case is matched to a control randomly selected from those subjects who were at risk and unexposed

at the case's failure time. Similarly, each unexposed case is matched to an exposed control. Analysis of the sampled data is performed using (2) with $w_i(t_j)$ the number of exposed at risk at the failure time if subject i is exposed at the failure time or the number of unexposed if i is unexposed.

An appealing feature of this method is that the full cohort information about the sampling stratification variable is summarized into the sample in the following sense. If the only covariates in the model are functions of the $A_j(t)$, it may be easily shown that (2) is the full cohort partial likelihood. Also, we note that there is no requirement that the sampling strata be included in the model as covariates for model comparison statistics, such as the likelihood ratio test between nested models, to be valid.

2 Notation and specification of the model

In this section we define a model for stratified nested case-control sampling. We fix throughout a time interval $[0,\tau]$, and following the counting process formulation of the Cox model as given by Andersen and Gill (1982), we let N_i, Y_i, and Z_i be the counting, censoring, and covariate processes for the ith subject, $i = 1,\ldots,n$. Moreover, we let A_i, which may be a function of Z_i, be the sampling stratum indicators with $A_i(t) \in \{1,\ldots,L\}$. At time t the risk set is $\mathcal{R}(t) = \{i : Y_i(t) = 1\}$, and the number of individuals at risk is $n(t) = |\mathcal{R}(t)| = \sum Y_i(t)$. As is usual, we assume that there is a non-decreasing family of σ-algebras $(\mathcal{H}_t)_{t\in[0,\tau]}$ such that the N_i are adapted to $(\mathcal{H}_t)$ and the Y_i, Z_i, and A_i are predictable with respect to $(\mathcal{H}_t)$. $\mathcal{H}_t$ is the "cohort history" including failure time, censoring, and covariate information up to time t.

The $(\mathcal{H}_t)$-intensity process λ_i of N_i is given heuristically by

$$\lambda_i(t)\mathrm{d}t = \mathrm{pr}\{\mathrm{d}N_i(t) = 1 \mid \mathcal{H}_{t-}\}, \tag{3}$$

where $\mathrm{d}N_i(t) = N_i\{(t+\mathrm{d}t)-\} - N_i(t-)$ is the increment of N_i over the small time interval $[t, t+\mathrm{d}t)$. Assuming censoring to be independent λ_i is by (1) given by

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta_0' Z_i(t)\}. \tag{4}$$

We write $\mathcal{R}_l(t) = \{i : Y_i(t) = 1, A_i(t) = l\}$ for the subset of $\mathcal{R}(t)$ which belongs to stratum l, and let $n_l(t) = |\mathcal{R}_l(t)|$ be the number of individuals in this stratum at time t. Then if a subject, say i, fails at time t, m_l controls are randomly sampled without replacement from $\mathcal{R}_l(t)$ except for the failure's stratum $\mathcal{R}_{A_i(t)}(t)$ from which $m_{A_i(t)} - 1$ are sampled from the $n_{A_i(t)}(t) - 1$ non-failures. We let $\tilde{\mathcal{R}}(t)$ denote the sampled risk set at t were a failure to occur at that time. This will consist of the failing individual together with its sampled set of controls. As a technical point, the number of controls could also depend on time. Specifically, if $n_l(t) < m_l$ for some l we would sample all individuals in this stratum but, for simplicity of exposition, we will assume below that the numbers of subjects sampled from each stratum do not depend on t. We introduce

$$\mathcal{P}(t) = \{\mathbf{r} \subset \mathcal{R}(t) : |\mathbf{r} \bigcap \mathcal{R}_l(t)| = m_l, l = 1,\ldots,L\},$$

and $\mathcal{P}_i(t) = \{\mathbf{r} \in \mathcal{P}(t) : i \in \mathbf{r}\}$. Then $\mathcal{P}(t)$ is the collection of all possible sets that may actually be used as sampled risk sets were a failure to occur at time t, while $\mathcal{P}_i(t)$ is the col-

lection of all possible sets if subject i is the failure. Note that there are $C(t)m_{A_i(t)}/n_{A_i(t)}(t)$ sets in $\mathcal{P}_i(t)$ where

$$C(t) = \prod_{l=1}^{L} \binom{n_l(t)}{m_l}.$$

We let $\mathcal{F}_{t-}$ contain information about all observed events in the cohort as well as about the sampling of controls in $[0,t)$. Thus $\mathcal{F}_{t-}$ is $\mathcal{H}_{t-}$ augmented with the sampling information. Then we have

$$\mathrm{pr}\{\tilde{\mathcal{R}}(t) = \mathbf{r} \mid \Delta N_i(t) = 1, \mathcal{F}_{t-}\} = C(t)^{-1} w_i(t) I\{\mathbf{r} \in \mathcal{P}_i(t)\}, \tag{5}$$

where $\Delta N_i(t) = N_i(t) - N_i(t-)$ is the increment of N_i at t, and $w_i(t) = n_{A_i(t)}(t)/m_{A_i(t)}$.

For each set $\mathbf{r} \in \mathcal{P}^{(m)}$, where $\mathcal{P}^{(m)}$ is the set of all subsets of $\{1,2,\ldots,n\}$ of size $m = \sum m_l$, we define $N_{(i,\mathbf{r})}(t)$ as the number of times in $[0,t]$ the ith individual fails and, at the same time, the sampled risk set equals $\mathbf{r}$. Moreover, we assume that the sampling is independent in the sense that the additional knowledge of which individuals have been sampled as controls before any time t do not alter the intensities of failures at t. Thus $\mathrm{pr}\{\mathrm{d}N_i(t) = 1 \mid \mathcal{F}_{t-}\} = \mathrm{pr}\{\mathrm{d}N_i(t) = 1 \mid \mathcal{H}_{t-}\}$. Informally therefore, by (3) and (5), the intensity process $\lambda_{(i,\mathbf{r})}$ of the counting process $N_{(i,\mathbf{r})}$ is given by

$$\begin{aligned}
\lambda_{(i,\mathbf{r})}(t)\mathrm{d}t &= \mathrm{pr}\{\mathrm{d}N_{(i,\mathbf{r})}(t) = 1 \mid \mathcal{F}_{t-}\} = \mathrm{pr}\{\mathrm{d}N_i(t) = 1, \tilde{\mathcal{R}}(t) = \mathbf{r} \mid \mathcal{F}_{t-}\} \\
&= \mathrm{pr}\{\mathrm{d}N_i(t) = 1 \mid \mathcal{F}_{t-}\} \times \mathrm{pr}\{\tilde{\mathcal{R}}(t) = \mathbf{r} \mid \Delta N_i(t) = 1, \mathcal{F}_{t-}\} \\
&= \lambda_i(t)\mathrm{d}t\, C(t)^{-1} w_i(t) I\{\mathbf{r} \in \mathcal{P}_i(t)\}.
\end{aligned}$$

These heuristics, combined with (4), imply that the counting processes $N_{(i,\mathbf{r})}$, for $\mathbf{r} \in \mathcal{P}^{(m)}$ and $i \in \mathbf{r}$, have intensity processes

$$\lambda_{(i,\mathbf{r})}(t) = Y_i(t)\lambda_0(t)\exp\{\beta_0' Z_i(t)\} C(t)^{-1} w_i(t) I\{\mathbf{r} \in \mathcal{P}_i(t)\}. \tag{6}$$

A formal treatment is given by Borgan *et al.* (1992). By standard counting process theory (e.g. Andersen *et al.*, 1992, Section II.4.1) it then follows that the

$$M_{(i,\mathbf{r})}(t) = N_{(i,\mathbf{r})}(t) - \int_0^t \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u \tag{7}$$

are local square integrable martingales. Their predictable variation processes are given as

$$\langle M_{(i,\mathbf{r})} \rangle(t) = \int_0^t \lambda_{(i,\mathbf{r})}(u)\mathrm{d}u, \tag{8}$$

while their predictable covariation processes are

$$\langle M_{(i,\mathbf{r})}, M_{(j,\mathbf{s})} \rangle(t) = 0 \tag{9}$$

for $(i,\mathbf{r}) \neq (j,\mathbf{s})$.

3 The partial likelihood and its basic likelihood properties

To derive the partial likelihood (2), we first introduce

$$N_{\mathbf{r}}(t) = \sum_{i \in \mathbf{r}} N_{(i,\mathbf{r})}(t) \tag{10}$$

for the process counting the number of times the sampled risk set equals $\mathbf{r}$ in $[0,t]$, and note that its intensity process, by (6), is

$$\lambda_{\mathbf{r}}(t) = \sum_{j \in \mathbf{r}} \lambda_0(t) \exp\{\beta_0' Z_j(t)\} C(t)^{-1} w_j(t) I\{\mathbf{r} \in \mathcal{P}_j(t)\}. \tag{11}$$

Then we factorise the intensity processes $\lambda_{(i,\mathbf{r})}$, not as just above (6), but as

$$\lambda_{(i,\mathbf{r})}(t) = \lambda_{\mathbf{r}}(t) \pi_t(i \,|\, \mathbf{r}),$$

where, by (6) and (11),

$$\pi_t(i \,|\, \mathbf{r}) = \frac{\exp\{\beta_0' Z_i(t)\} w_i(t)}{\sum_{j \in \mathbf{r}} \exp\{\beta_0' Z_j(t)\} w_j(t)}. \tag{12}$$

Note that (12) has the interpretation of the conditional probability of the ith individual failing at t, given $\mathcal{F}_{t-}$ and that there is a failure among individuals in the set $\mathbf{r}$ at t. The partial likelihood (2) is then obtained by multiplying together conditional probabilities of the form (12) for each failure time and sampled risk set.

We will show that (2) has basic likelihood properties, i.e. that the score vector has expectation zero, and that its covariance matrix equals the expected information matrix. To this end we note that (2) may may be written as $\mathcal{L}(\beta, \tau)$, where

$$\mathcal{L}(\beta, t) = \prod_{u \in [0,t]} \prod_{\mathbf{r} \in \mathcal{P}(u)} \prod_{i \in \mathbf{r}} \left[\frac{\exp\{\beta' Z_i(u)\} w_i(u)}{\sum_{j \in \mathbf{r}} \exp\{\beta' Z_j(u)\} w_j(u)} \right]^{\Delta N_{(i,\mathbf{r})}(u)}.$$

We introduce the notation

$$S_{\mathbf{r}}^{(\gamma)}(\beta, t) = \sum_{j \in \mathbf{r}} Z_j(t)^{\otimes \gamma} \exp\{\beta' Z_j(t)\} w_j(t) \tag{13}$$

for $\gamma = 0, 1, 2$, where for a vector a, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa'$, and let

$$E_{\mathbf{r}}(\beta, t) = S_{\mathbf{r}}^{(1)}(\beta, t) / S_{\mathbf{r}}^{(0)}(\beta, t) \tag{14}$$

and

$$V_{\mathbf{r}}(\beta, t) = \frac{S_{\mathbf{r}}^{(2)}(\beta, t)}{S_{\mathbf{r}}^{(0)}(\beta, t)} - E_{\mathbf{r}}(\beta, t)^{\otimes 2}. \tag{15}$$

Then, apart from a constant term,

$$\log \mathcal{L}(\beta, t) = \int_0^t \sum_{\mathbf{r} \in \mathcal{P}(u)} \sum_{i \in \mathbf{r}} \left[\beta' Z_i(u) - \log\{S_{\mathbf{r}}^{(0)}(\beta, u)\} \right] \mathrm{d}N_{(i,\mathbf{r})}(u).$$

Differentiation with respect to β in the usual way then yields the "score vector process"

$$U(\beta,t) = \frac{\partial}{\partial\beta}\log\mathcal{L}(\beta,t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)}\sum_{i\in\mathbf{r}} \{Z_i(u) - E_{\mathbf{r}}(\beta,u)\}\,\mathrm{d}N_{(i,\mathbf{r})}(u), \tag{16}$$

and the "information matrix process"

$$\mathcal{I}(\beta,t) = -\frac{\partial^2}{\partial\beta^2}\log\mathcal{L}(\beta,t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)} V_{\mathbf{r}}(\beta,u)\mathrm{d}N_{\mathbf{r}}(u). \tag{17}$$

Using (6), (7), (13) and (14) it is seen that the score process, evaluated at the true parameter vector β_0, equals

$$U(\beta_0,t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)}\sum_{i\in\mathbf{r}} \{Z_i(u) - E_{\mathbf{r}}(\beta_0,u)\}\,\mathrm{d}M_{(i,\mathbf{r})}(u), \tag{18}$$

i.e. it is a vector valued stochastic integral, and therefore a local square integrable martingale. In particular, the expected score is zero provided that the expectation exists.

The predictable variation process of (18) is, by (6)-(9) and (13)-(15),

$$\begin{aligned}\langle U(\beta_0,\cdot)\rangle(t) &= \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)}\sum_{i\in\mathbf{r}} \{Z_i(u) - E_{\mathbf{r}}(\beta_0,u)\}^{\otimes 2}\,\lambda_{(i,\mathbf{r})}(u)\mathrm{d}u \\ &= \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)} V_{\mathbf{r}}(\beta_0,u)S_{\mathbf{r}}^{(0)}(\beta_0,u)C(u)^{-1}\lambda_0(u)\mathrm{d}u. \end{aligned}\tag{19}$$

Moreover, (17) evaluated at β_0 may be written as

$$\mathcal{I}(\beta_0,t) = \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)} V_{\mathbf{r}}(\beta_0,u)\lambda_{\mathbf{r}}(u)\mathrm{d}u + \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)} V_{\mathbf{r}}(\beta_0,u)\mathrm{d}M_{\mathbf{r}}(u),$$

where, by (10) and (11), the

$$M_{\mathbf{r}}(t) = N_{\mathbf{r}}(t) - \int_0^t \lambda_{\mathbf{r}}(u)\mathrm{d}u$$

are local square integrable martingales. Therefore, by (11), (13) and (19),

$$\mathcal{I}(\beta_0,t) = \langle U(\beta_0,\cdot)\rangle(t) + \int_0^t \sum_{\mathbf{r}\in\mathcal{P}(u)} V_{\mathbf{r}}(\beta_0,u)\mathrm{d}M_{\mathbf{r}}(u).$$

Thus the observed information equals the predictable variation of the score plus a local square integrable martingale. In particular, by taking expectations, assuming that they exist, it follows that the expected information matrix equals the covariance matrix of the score.

Let $\widehat{\beta}$ be the maximum partial likelihood estimator obtained by maximizing (2). From the above results it then follows by a Taylor series expansion in the usual way (see Borgan *et al.*, 1992, for details) that, under appropriate conditions, $\sqrt{n}\left(\widehat{\beta} - \beta_0\right)$ converges in distribution to a multivariate normal random variable with mean zero and covariance matrix Σ^{-1} described in Borgan *et al.* (1992). Moreover the asymptotic information matrix Σ may be obtained as the limit in probability of $1/n$ times (19).

4 Asymptotic relative efficiencies

This section presents a large sample comparison of simple and stratified nested case-control sampling when there is a binary "exposure" variable Z_1, assumed to be known for the full cohort and which will serve as the sampling stratification variable, and a binary "confounder" variable Z_2 to be collected for the sampled subjects. For the simple nested case-control design, sampled risk sets of size m consist of the failure and $m-1$ controls randomly sampled without regard to exposure status. For the stratified sample, m_0 unexposed, with $Z_1 = 0$, and m_1 exposed, with $Z_1 = 1$, subjects are sampled, with $m_0 + m_1 = m$. In a model analogous to that considered by Breslow and Cain (1988) for two stage sampling for unconditional logistic models, let the intensity process for an individual with covariates Z_1 and Z_2 be specified as

$$\lambda(t) = Y(t)\lambda_0(t)\exp(Z_1\beta_1 + Z_2\beta_2), \tag{20}$$

c.f. (4).

We assume that the joint exposure-confounder probability distribution for those at risk remain constant over time. Subjects in the cohort are assumed to arise as independent and identically distributed realizations from the covariate and censoring distributions with failure time distributions determined by the associated hazard functions. We note that, under this assumption, asymptotic relative efficiencies do not depend on the failure rate (Goldstein and Langholz, 1992, Example 6.3b). We are interested in comparing the variances of $\widehat{\beta}_1$ after controlling for the effect of Z_2 for the two sampling designs. It is also of interest to compare the variances of $\widehat{\beta}_3$ for a model with an interaction term $Z_1Z_2\beta_3$ added in (20), assuming that $\beta_3 = 0$. The strategy we used was to compute the 3×3 asymptotic information matrix Σ for the interaction model. The variance of $\widehat{\beta}_1$ controlling for Z_2 in the model (20) is then obtained as the $(1,1)$th entry after inverting the upper left 2×2 part of Σ. The variance of $\widehat{\beta}_3$ is given as the $(3,3)$th entry in Σ^{-1}. The general formulas for asymptotic variance for the two designs are found in Borgan *et al.* (1992, Section 7). Confounding between Z_1 and Z_2 was measured by the population odds ratio

$$\theta = \frac{\mathrm{pr}(Z_1 = 1, Z_2 = 1)\mathrm{pr}(Z_1 = 0, Z_2 = 0)}{\mathrm{pr}(Z_1 = 1, Z_2 = 0)\mathrm{pr}(Z_1 = 0, Z_2 = 1)}.$$

Table 1 gives asymptotic relative efficiencies of the "balanced" stratified design, i.e., $m_0 = m_1$, relative to simple nested case-control sampling. These are given for $\mathrm{pr}(Z_1(t) = 1) \equiv 0.05$ and $\mathrm{pr}(Z_2(t) = 1) \equiv 0.3$ and various values of m, $\exp(\beta_1)$, $\exp(\beta_2)$ and θ. In every case, stratification results in substantial gains in efficiency for estimating β_1 after controlling for Z_2, and for the estimation of β_3. For 1:1 matching, there are large losses of efficiency for estimating β_2 after controlling for Z_1 but, since Z_2 is a confounder in the model, a precise estimate of its effect is not important. We also computed the asymptotic relative efficiencies of the balanced design compared to the stratified sample with optimal m_0 and m_1 for $m = 4$ or 8. For all combinations of parameters used in Table 1, the there was very little difference between the balanced and optimal designs.

We also considered the situation where the observed relative risk for exposure is explained by the confounder. In this situation, $\beta_1 = 0$ and β_2 was chosen to yield a a marginal relative risk for exposure of $\exp(\beta_1^*) = 1.5$ or 2 when full cohort data is used to estimate β_1 without Z_2 in the model. The results for balanced stratified sampling are given in Table 2.

Table 1: Asymptotic relative efficiencies of stratified versus simple nested case-control sampling when the exposure relative risk $\exp(\beta_1)$ is 2 or 4 [a].

		(a) $\exp(\beta_1) = 2$					
		1:1 matching ($m = 2$)			1:3 matching ($m = 4$)		
e^{β_2}	θ	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
0.2	0.2	2.65	0.45	1.56	1.73	0.83	1.17
0.2	0.5	2.75	0.45	1.54	1.72	0.84	1.17
0.2	1.0	2.83	0.45	1.53	1.71	0.85	1.18
0.2	2.0	2.86	0.45	1.53	1.68	0.86	1.20
0.2	5.0	2.63	0.43	1.51	1.59	0.87	1.22
1.0	0.2	2.46	0.25	2.35	1.58	0.75	1.54
1.0	0.5	2.75	0.30	2.24	1.60	0.77	1.52
1.0	1.0	2.86	0.35	2.16	1.61	0.79	1.50
1.0	2.0	2.73	0.38	2.04	1.60	0.80	1.48
1.0	5.0	2.23	0.37	1.76	1.54	0.80	1.43
5.0	0.2	2.56	0.26	3.30	1.55	0.67	2.02
5.0	0.5	2.85	0.35	2.95	1.71	0.71	1.92
5.0	1.0	2.80	0.42	2.61	1.80	0.73	1.83
5.0	2.0	2.58	0.46	2.19	1.82	0.74	1.71
5.0	5.0	2.24	0.49	1.66	1.78	0.74	1.51
		(b) $\exp(\beta_1) = 4$					
		1:1 matching ($m = 2$)			1:3 matching ($m = 4$)		
e^{β_2}	θ	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
0.2	0.2	4.06	0.70	1.97	2.23	0.90	1.31
0.2	0.5	4.15	0.71	1.97	2.23	0.92	1.32
0.2	1.0	4.20	0.71	1.98	2.21	0.94	1.33
0.2	2.0	4.13	0.72	2.00	2.16	0.96	1.36
0.2	5.0	3.63	0.70	1.99	2.02	0.99	1.40
1.0	0.2	3.75	0.44	3.42	2.01	0.82	1.94
1.0	0.5	4.19	0.54	3.25	2.07	0.87	1.90
1.0	1.0	4.35	0.62	3.13	2.09	0.90	1.86
1.0	2.0	4.15	0.67	2.95	2.06	0.93	1.80
1.0	5.0	3.41	0.65	2.52	1.93	0.92	1.69
5.0	0.2	3.46	0.43	4.62	1.89	0.77	2.66
5.0	0.5	3.99	0.58	4.12	2.14	0.84	2.45
5.0	1.0	4.07	0.69	3.61	2.26	0.87	2.24
5.0	2.0	3.88	0.76	2.94	2.29	0.89	1.95
5.0	5.0	3.47	0.80	2.02	2.23	0.89	1.51

[a] For $m_0 = m_1$ and $\text{pr}(Z_1 = 1) = 0.05$ and $\text{pr}(Z_2 = 1) = 0.30$. Efficiencies for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ based on fitting the model with no interaction term '$Z_1 Z_2$'.

Table 2: Asymptotic relative efficiencies of stratified versus simple nested case-control sampling when there is a marginal relative risk for exposure $\exp(\beta_1^*)$ of 1.5 or 2.0 but $\exp(\beta_1) = 1$ after controlling for confounding[a].

		(a) $\exp(\beta_1^*) = 1.5$					
		1:1 matching ($m = 2$)			1:3 matching ($m = 4$)		
e^{β_2}	θ	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
36.7	2.0	1.91	0.62	1.18	1.55	0.77	1.01
3.7	4.0	1.54	0.25	1.35	1.45	0.66	1.40
2.8	6.0	1.39	0.22	1.22	1.40	0.67	1.35
2.5	8.0	1.28	0.21	1.13	1.38	0.67	1.33
2.3	10.0	1.21	0.20	1.07	1.36	0.67	1.31
		(b) $\exp(\beta_1^*) = 2$					
		1:1 matching ($m = 2$)			1:3 matching ($m = 4$)		
e^{β_2}	θ	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
22.8	4.0	1.79	0.60	1.03	1.53	0.76	1.04
8.1	6.0	1.54	0.35	1.18	1.51	0.66	1.33
6.0	8.0	1.42	0.30	1.14	1.48	0.66	1.34
5.1	10.0	1.35	0.28	1.09	1.46	0.65	1.33

[a] See footnote to Table 1.

Once again stratified sampling offers considerable efficiency advantage for estimation of the exposure effect after controlling for the confounder.

These asymptotic relative efficiency calculations compare the two designs with comparable matching ratios. However, the number of distinct subjects in the sample, often the relevant determinant of the cost of the sampled study (Langholz and Thomas, 1991), is different for the two designs. If this is the case, the relative efficiencies underestimate the advantage of the stratified design since the stratified design requires a smaller proportion of the cohort than does the standard. This is because in the stratified design the exposed subjects have a higher chance of appearing in multiple sampled risk sets than with the simple design. The difference in the proportion sampled increases with decreasing probability of exposure, increasing disease probability, and increasing matching ratio. Thus, if the relevant measure of sample size is the number of distinct subjects, the actual relative efficiencies may be substantually higher than suggested in the tables.

5 Non-representative sampling of cases

If the disease of interest is rare, as is often the situation in epidemiologic studies, all cases will be used in the sampled data set. But, occasionally, especially if there are many cases available and few of them are "exposed," some type of non-representative sampling of the cases may be desirable. This, of course, would be done before sampling controls since sampled risk sets are uninformative about β_0 if there is no case. Stratified sampling in which a fixed number of cases is picked from each sampling stratum, as we have described

for the controls, cannot be accomodated by our methodology because the total number of cases in each sampling stratum for the entire study period is not in $\mathcal{F}_{t-}$ for $t < \tau$. We propose an approximation to this type of sampling in which the decision to include a case is determined by a Bernulli trial with probability depending on the sampling stratum of the case and chosen to yield the desired proportions of cases from each stratum. A stratified sample of the controls is picked as before for each included case.

A simple generalization of the hazard model (6) is required to accomodate this type of sampling. If a failure were to occur at time t, let $B(t)$ indicate whether that failure would be included in the sample and $p_i(t)$ be the probability of inclusion if subject i was that failure. Typically, the $p_i(t)$ will take on a small number of values depending upon subject i's sampling stratum but, in the development that follows, it is not required. We now redefine $\mathcal{F}_{t-}$ to additionally include the case sampling information in $[0, t)$. Analogous to (5)

$$\begin{aligned}&\mathrm{pr}\{\tilde{\mathcal{R}}(t) = \mathbf{r}, B(t) = b \mid \Delta N_i(t) = 1, \mathcal{F}_{t-}\} = \\ &\quad \mathrm{pr}\{\tilde{\mathcal{R}}(t) = \mathbf{r} \mid \Delta N_i(t) = 1, B(t) = b, \mathcal{F}_{t-}\} \times \mathrm{pr}\{B(t) = b \mid \Delta N_i(t) = 1, \mathcal{F}_{t-}\}.\end{aligned}$$

The first probability is as in (5) if $b = 1$ and $I(\mathbf{r} = \{i\})$ if $b = 0$. The second probability is $p_i(t)$ or $1 - p_i(t)$ for $b = 0, 1$, respectively. We define $N_{(i,\mathbf{r},b)}(t)$, for $\mathbf{r} \in \mathcal{P}^{(m)} \cup \{i\}$, $i \in \mathbf{r}$, and $b = 0, 1$, as the number of times in $[0, t]$ a failure of the ith individual is not included, respectively included, in the sample and the sampled risk set equals $\mathbf{r}$. Then the counting processes $N_{(i,\mathbf{r},b)}$ have intensity processes

$$\begin{aligned}\lambda_{(i,\mathbf{r},b)}(t) \;=\; & Y_i(t)\lambda_0(t)\exp\{\beta_0' Z_i(t)\}C(t)^{-1}w_i(t)p_i(t)I\{\mathbf{r} \in \mathcal{P}_i(t), b = 1\} \\ & + Y_i(t)\lambda_0(t)\exp\{\beta_0' Z_i(t)\}\{1 - p_i(t)\}I(\mathbf{r} = \{i\}, b = 0).\end{aligned}$$

The partial likelihood may be derived as in (10) - (12) with

$$\pi_t(i|\mathbf{r}, 1) = \frac{\exp\{\beta_0' Z_i(t)\}w_i(t)p_i(t)}{\sum_{j\in\mathbf{r}}\exp\{\beta_0' Z_j(t)\}w_j(t)p_j(t)}$$

if $\mathbf{r} \in \mathcal{P}_i(t)$ and $\pi_t(i|\mathbf{r}, 0) = 1$ if $\mathbf{r} = \{i\}$.

Thus, letting $\tilde{F}$ denote the set of failure times for the sampled failures, the partial likelihood for the data may be written in standard notation as

$$\mathcal{L}(\beta) = \prod_{t_j\in\tilde{F}}\left[\frac{\exp\{\beta' Z_{i_j}(t_j)\}w_{i_j}(t_j)p_{i_j}(t_j)}{\sum_{k\in\tilde{\mathcal{R}}(t_j)}\exp\{\beta' Z_k(t_j)\}w_k(t_j)p_k(t_j)}\right]. \tag{21}$$

The basic likelihood properties are derived in a manner analogous to Section 3 using that

$$M_{(i,\mathbf{r},b)}(t) = N_{(i,\mathbf{r},b)}(t) - \int_0^t \lambda_{(i,\mathbf{r},b)}(u)\mathrm{d}u$$

are local square integrable martingales.

6 Discussion

We have developed stratified sampling assuming the simple Cox model (1). The methods easily generalize to accomodate multiple population strata with a different baseline hazard for each stratum. In this case, stratifed sampling would be performed within population stratum. Also, the exponential form of the relative risk may be replaced by $r\{\beta_0' Z(t)\}$ for a general relative risk function $r(\cdot)$ with $r(0) = 1$. Further, the expressions apply without change to multiple event data.

The stratified sampling itself is more flexible than may be interpreted by our presentation. We have discussed stratification based on a subject's absolute exposure measure. Another valid method is to define sampling strata based on quantiles of the exposure values of at risk subjects. Stratification might also be based on determinants of cost of collecting data. For example, in an occupational cohort of a particular factory, many cohort members may have ceased employment at the factory during the follow up period. The sampling strata could be based on employment status and designed to over-sample those who are still employed since it is easy to contact them. This cannot be done based on present employment status but must be based on employment status at the time of disease being considered so the benefit of this strategy will depend upon how recently the cases of disease generally occurred.

It may well be the situation that the sampling stratum indicator A summarizes a continuous or multilevel covariate Z_1, which is available on the full cohort, into L sampling strata categories. As mentioned in Section 1, models fitted using the stratified sample which are based just on functions of A retain the full cohort information about A. However, since they are more precise, one would typically use the actual values of $Z_{i1}(t_j)$ when analyzing the stratified sampled data, and the information will be less than that of the full cohort because of the grouping used to define the sampling strata. The question then arises of how to form the sampling strata groupings so as to retain as much information as possible about Z_1 given that the other covariates are not known. Let $(c_{l-1}(t_j), c_l(t_j)]$ be non-overlapping intervals with $c_0(t_j) = -\infty$, $c_L(t_j) = \infty$ and the sampling strata be defined by $\mathcal{R}_l(t_j) = \{i \in \mathcal{R}(t_j) : c_{l-1}(t_j) < Z_{i1}(t_j) \leq c_l(t_j)\}$. We conjecture that a good strategy is to set $L = m$ and choose the $c_l(t_j)$ so that the conditional probability of disease, based only on Z_1, is $1/m$ in each interval. This depends, of course, on how Z_1 is modelled and, assuming for example a trend model, may be approximated by choosing the $c_l(t_j)$ such that

$$\frac{\sum_{i \in \mathcal{R}(t_j)} \exp\{\widehat{\beta} Z_{i1}(t_j)\} I\{c_{l-1}(t_j) < Z_{i1}(t_j) \leq c_l(t_j)\}}{\sum_{i \in \mathcal{R}(t_j)} \exp\{\widehat{\beta} Z_{i1}(t_j)\}} \approx 1/m.$$

This strategy did appear to be best in a preliminary empirical investigation of the $m = 2$ situation but further work is needed to determine if it generally near optimal.

We have assumed that sampling strata are known for all cohort subjects at the time when the sampling is performed. In many situations, the sampling strata may be available for a portion of the cohort but missing for the rest. Suppose, for ease of exposition, that the sampling strata are based on a dichotomous exposure covariate with the probability of exposure small and that most of the missing values could be filled in for the sampled data. If the number of subjects missing exposure data is small, they could be included with the unexposed forming an "unexposed or missing information" sampling stratum. If there is a large number of such subjects, they could form another sampling stratum and a design with three strata could be considered. In either situation, exposure status would

be ascertained for those sampled individuals with missing exposure and this would be used in the analysis of the sampled data. The weight associated with such subjects would not depend on their final exposure status but would reflect the sampling stratum from which they were picked.

Stratified sampling may also be possible when it is not possible to identify all subjects at risk but it is only possible to insure that controls can be randomly sampled from the sampling strata that make up the risk set at a given failure time. This would be the situation in a stratified population based age-matched case-control study. Note that the $n_l(t_j)$ may be replaced by $\pi_l(t_j) = n_l(t_j)/n(t_j)$ or by $\pi_l(t_j)/\pi_1(t_j)$ in the weights without changing the partial likelihood (2). This suggests that if the proportion or ratio of proportions of the population in each sampling stratum is known as a function of time, these methods may be applied.

Finally, if the $\pi_l(t_j)$ are not known, they could be estimated from a sample of the population. This could be done in a number of ways. One is to do a survey of the population and estimate the $\pi_l(t_j)$ from this sample. Another is to use a two-stage sampling procedure, analogous to that of Breslow and Cain (1988). In this method a "first stage" sample, a random sample of potential "controls" without regard to sampling strata is picked for each case. The sampling stratum of each subject in the first stage sample is determined and the "second stage" stratified sample is then picked. The $\pi_l(t_j)$ would be estimated from the first stage sample. Further work is needed to assess the validity of these approaches and to develope variance adjustment methods to account for the additional variation resulting from the estimation of the weights.

Acknowledgements

This research was initiated when the both authors were on sabbatical leave at the MRC Biostatistics Unit, Cambridge, England, the academic year 1991/92. The MRC Biostatistics Unit is acknowledged for its hospitality and for providing us with the best working facilities during this year. The authors wish to thank David Clayton of the Unit for pointing out the importance of this problem. We gratefully acknowledge the support of this work by the Norwegian Research Council for Science and the Humanities and the United States National Cancer Institute.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1992). *Statistical Models Based on Counting Processes.* Springer Verlag, New York. (in press).

Borgan, Ø., Langholz, B., and Goldstein, L. (1992). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. Statistical research report, Institute of Mathematics, University of Oslo.

Breslow, N. and Cain, K. (1988). Logistic regression for two stage case-control data. *Biometrika* **75**, 11–20.

Cain, K. and Breslow, N. (1988). Logistic regression analysis and efficient design for two stage studies. *Am. J. Epidemiol.* **128**, 1198–206.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20**. (in press).

Langholz, B. and Thomas, D. C. (1991). Efficiency of cohort sampling designs: Some surprising results. *Biometrics* **47**, 1563–71.

Oakes, D. (1981). Survival times: Aspects of partial likelihood (with discussion). *Internat. Statist. Rev.* **49**, 235–64.

Thomas, D. C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. By F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *J. Roy. Statist. Soc. A* **140**, 469–91.

Weinberg, C. and Sandler, D. (1991). Randomized recruitment in case-control studies. *Am. J. Epidemiol.* **134**, 421–32.

Weinberg, C. and Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling. *Biometrics*, **46** 963–75.

White, J. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. of Epidemiol.* **115**, 119–28.